# REFERENCE-AWARE FAIRNESS AUDITING: SEPARATING LABEL ARTIFACTS FROM MODEL BIAS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Fairness audits compare model errors across subgroups, but rely on the assumption that evaluation labels are equally accurate for every subgroup. When label noise differs across subgroups, part of the reported disparity can come from the labels rather than the model. We show that the two can be separated when label errors are not systematically related to the model's own errors, and propose *Reference-Aware Fairness Auditing* (RAFA): estimate the label noise per subgroup, subtract its variance from the measured error, and report the adjusted disparity alongside the conventional one. In simulation, unequal label noise alone makes a fair model appear unfair, and RAFA removes $93\%$ of the label-induced disparity. In controlled experiments on real lung-nodule and chest X-ray predictions, we inject label-noise and model-error disparities of known size. RAFA removes the label-induced part and preserves the known model part: adjusted and known model gaps differ by $0.003$–$0.004$ on the nodules and below $0.002$ on the X-rays, while the injected gaps reach $0.9$ and $0.6$. We then apply the full procedure to fetal-weight estimation from ultrasound ($31{,}386$ examinations, 17 hospitals), where the evaluation label is birth weight projected back to scan time, so its noise grows with the scan-to-delivery interval, as a growth model predicts. RAFA attributes $73\%$ of the measured disparity to unequal label noise (bootstrap median $64\%$, $95\%$ CI $37$–$83\%$). Finally, we give three explicit conditions that determine when RAFA applies and show how to check them in practice, using committee-labeled nodules and crowd counting as diagnostic examples.

## 1 INTRODUCTION

Fairness auditing of predictive models is becoming increasingly common: model performance is compared across subgroups, and observed disparities may indicate that a model serves some groups less well than others (Seyyed-Kalantari et al., 2021; Gichoya et al., 2022; Barocas et al., 2023). Such comparisons typically assume that the evaluation labels have similar reliablility across subgroups. When label reliability differs, however, measured subgroup differences may reflect not only model error, but also differences in label quality. This can make it difficult to determine whether an observed disparity originates from the model or from the evaluation process itself.

Evaluation labels are not always error-free, and their accuracy can depend on how they are generated or measured (Frénay & Verleysen, 2014; Mishra & Gorana, 2021). This is especially important for fairness auditing when label noise differs across subgroups. Prior work has shown that group-dependent label noise can distort both fairness evaluation and mitigation (Fogliato et al., 2020; Wang et al., 2020b; Adebayo et al., 2023; Mhasawade et al., 2024; Akintande et al., 2025), while related studies have examined uncertainty and label bias in fairness-sensitive medical imaging (Tahir et al., 2023; Kuzucu et al., 2023; Ali et al., 2021; Jones et al., 2024; Parikh et al., 2025; 2026; Stanley et al., 2026). Existing approaches often address this problem by correcting labels or modifying training (Li et al., 2023; Sengamedu & Pham, 2023). In contrast, RAFA leaves the trained model unchanged and treats the audit itself as a measurement-error problem (Fuller, 1987; Taylor, 1997): an observed subgroup gap may reflect both label quality and model behavior. Figure 1 illustrates this distinction and the central idea of RAFA: separating label-induced disparity from model disparity.

We organize the paper around four questions: (i) Can unequal label noise create an apparent fairness gap? (ii) Can RAFA remove the label-induced component while preserving genuine model disparity?



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

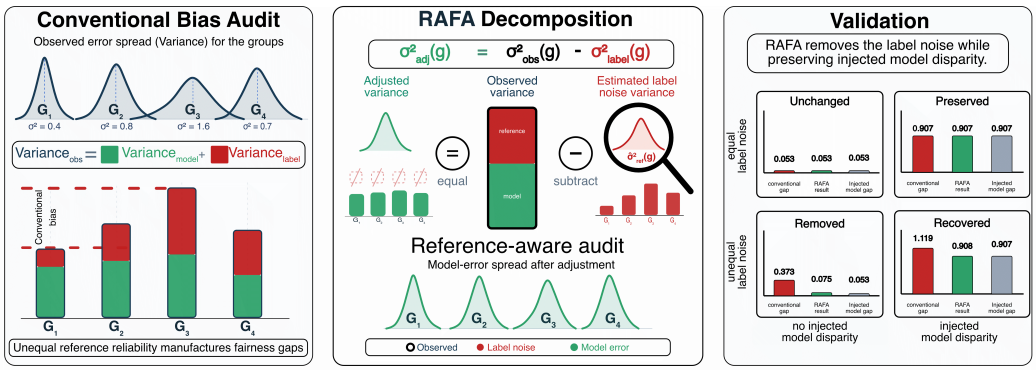



Figure 1: **Unequal label noise across subgroups can create apparent fairness gaps. Left:** a conventional audit can report a larger subgroup error simply because its evaluation labels are noisier. **Middle:** RAFA uses an estimate of label noise to report a label-noise-adjusted disparity. **Right:** in controlled experiments on real lung-nodule predictions, RAFA removes injected label-noise disparity while preserving injected model disparity (Table 1).

(iii) Can the required label-noise estimate be obtained in a real application? and (iv) under what conditions is RAFA valid? Our contributions map onto these:

- **Formalization (Section 2).** We show that observed subgroup error mixes model error with evaluation-label noise, and formalize when their variance contributions can be separated. RAFA uses an application-specific estimate of subgroup label noise to report a label-noise-adjusted disparity.
- **Controlled validation (Sections 3).** Simulation shows that unequal label noise alone can create an apparent fairness gap. Controlled injections into real lung-nodule and chest-radiograph predictions show that RAFA removes the label-induced component while preserving known model disparity, including on real demographic subgroups.
- **Real-world application (Section 4).** We apply the full RAFA procedure to fetal-weight estimation, where the evaluation label is birth weight projected back to scan time. We estimate the label noise from the data and find that unequal label noise explains most of the measured disparity.
- **Scope (Section 5).** We state three conditions, R1–R3, for when RAFA is valid, and document settings that violate them, from committee-labeled nodules to crowd counting.

## 2 Reference-Aware Fairness Auditing

### 2.1 Separating model error from label noise

A fairness audit measures model error against an evaluation label, but the label itself can be wrong. The audit therefore sees two error sources mixed together: the model's error and the label's error. RAFA aims to separate them.

Let $\hat{y}$ be the model prediction and $y_{true}$ the unobservable true target. The observed label used for evaluation is

$$y_{obs} = y_{true} + \epsilon, \tag{1}$$

where $\epsilon$ is label noise with $\mathbb{E}[\epsilon \mid g] = 0$: Labels may be noisier in some subgroups $g$, but within a subgroup they are not systematically shifted above or below the truth, in other words, we do not have label bias, only label noise. A bias audit performed with respect to the $y_{obs}$ observes the signed error

$$e \;=\; \hat{y} - y_{obs} \;=\; \underbrace{(\hat{y} - y_{true})}_{\text{model error } e_m} - \epsilon. \tag{2}$$

Within subgroup $g$, writing $\sigma^2_{\text{model}}(g) := \text{Var}(e_m \mid g)$ and $\sigma^2_{\text{label}}(g) := \text{Var}(\epsilon \mid g)$,

$$\text{Var}(e \mid g) \;=\; \sigma^2_{\text{model}}(g) + \sigma^2_{\text{label}}(g) - 2\,\text{Cov}(e_m, \epsilon \mid g). \tag{3}$$

Two subgroups with the same model error can therefore show different measured error, purely because their labels differ in noise.

**Assumption.** RAFA assumes zero covariance between model error and label noise within each subgroup,

$$\text{Cov}(e_m, \epsilon \mid g) = 0. \tag{4}$$

Please note that this does not require full statistical independence. It simply means that the label's errors and the model's errors are not driven by the same cases. The value of this assumption is that under Eq. (4) the variances simply add:

$$\text{Var}(e \mid g) = \sigma^2_{\text{model}}(g) + \sigma^2_{\text{label}}(g). \tag{5}$$

**The adjustment.** If the subgroup label-noise variance can be estimated as $\hat{\sigma_{\text{label}}}^2(g)$, RAFA removes this contribution from the measured variance:

$$\hat{s}_{\text{adj}}(g) \;=\; \sqrt{\max\big(\text{Var}(e \mid g) - \hat{\sigma_{\text{label}}}^2(g),\, 0\big)}. \tag{6}$$

The estimate $\hat{\sigma_{\text{label}}}^2(g)$ is application-specific: repeated measurements, multiple annotators, known measurement uncertainty, or the physics of the labeling process. The clip at zero plays two roles. Small negative values arise from sampling noise even when the method is valid, and the clip keeps the estimate defined, at the cost of a small upward bias when the true adjusted error is near zero. Systematic clipping, where the noise estimate exceeds the observed variance in every subgroup, is instead a failure signal: the noise is over-estimated or the assumptions do not hold (Section 5). We subtract on the variance scale because independent error sources add exactly there; no such additivity holds for mean absolute error. For interpretability, we report errors as standard deviations, $\sqrt{\text{Var}}$.

**When is the assumption plausible?** Zero covariance is plausible when the model and the evaluation label obtain their information through largely separate processes. For example, a model may predict from an image while the evaluation label is derived from an independently measured outcome. The assumption is less plausible when the model and the labeling process depend on the same information. Shared factors such as poor image quality, ambiguous structure, or intrinsic task difficulty may then cause both to fail on the same examples. Section 5 gives a concrete example in crowd counting. Using out-of-sample predictions prevents memorization of the particular evaluation-label errors, but does not by itself guarantee Eq. (4). We treat zero covariance as an application-specific assumption that must be justified from how the model predictions and evaluation labels are generated.

**If the assumption fails.** From Eq. (3), subtracting $\sigma^2_{\text{label}}(g)$ leaves $\sigma^2_{\text{model}}(g) - 2\,\text{Cov}(e_m, \epsilon \mid g)$. Given a positive covariance, where the same cases are hard both for the labeler and the model, the adjustment will underestimate model error and may hide genuine disparity; this is the unsafe direction, and its practical signature is an implausibly small adjusted error, in the extreme the clip binding (Section 5). On the other hand, a negative covariance leaves the adjusted error above the true model error and is conservative. The covariance cannot be identified from the observed error alone, so the condition must be assessed from how the data are generated, not from $\text{Var}(e \mid g)$ itself.

### 2.2 THE RAFA PROTOCOL

Given model predictions, evaluation labels, and subgroups, RAFA follows the four-step procedure in Fig. 2. RAFA adjusts the *evaluation*, not the model: it requires no retraining and does not change model predictions.

**Requirements.** RAFA relies on three conditions. **R1:** label noise exists and varies across subgroups. **R2:** the estimation process for label noise does not depend on the audited model's errors: either it never uses them, or it uses them in a way where the model's own error cancels out of the estimate (Section 4). **R3:** within each subgroup, model error and label noise are uncorrelated (Eq. 4). R3 makes the subtraction valid; R2 makes the subtracted number trustworthy. Neither can be tested from the audit data alone, so we argue both from how the labels and the estimate are produced.




Figure 2: **RAFA protocol.** The estimated subgroup label noise is used to separate the label-induced component of the observed error before reporting conventional and adjusted disparities.

**Disparity and uncertainty.** We define disparity as the range of subgroup error, $\Delta = \max_g s(g) - \min_g s(g)$. The adjustment itself is made per subgroup, so any subgroup summary can be applied after it; we use the range because it captures the extremes an audit cares about and is common in practice. Confidence intervals are obtained by cluster bootstrap (Efron & Tibshirani, 1993), re-estimating the label-noise term within each replicate so that uncertainty in both the observed error and the noise estimate is propagated.

## 3 Controlled validation

We first check the mechanism in the simplest possible setting. There is no trained model here: we draw model errors and label noise directly from known distributions. Four subgroups receive identical model errors (SD $0.40$) but increasingly noisy labels (noise SD $0.30$, $0.45$, $0.58$, $0.70$). A conventional audit reports a disparity of $0.241$, although the true model gap is only $0.002$: the reported gap is almost entirely a label artifact, the mechanism of Figure 1 (left). Given the known noise levels, RAFA reduces the gap to $0.017$, removing $93\%$ of the label-induced disparity.

We now test RAFA on real model outputs, where we control exactly how much disparity comes from label noise and how much comes from the model. We use two datasets. On LIDC-IDRI (Armato III et al., 2011), a lung-nodule malignancy regressor is trained with patient-level cross-validation, and its out-of-fold predictions are fixed throughout (mean absolute error $0.70$ on the $1$–$5$ scale). On CheXpert (Irvin et al., 2019), each radiograph is labeled for a set of *findings*, clinical observations such as edema or consolidation, extracted automatically from the radiology report. A small validation set additionally carries labels from a panel of three radiologists, which we use to score the model's predictions. A DenseNet-121 classifier is trained with patient-level splits and its predictions frozen; we evaluate two findings, Edema and Consolidation, on $18{,}654$ held-out images (validation AUC $0.90$ and $0.91$ against the radiologist labels).

**Design.** We form subgroups and independently manipulate two quantities: **(i)** label-noise disparity, by adding mean-zero noise with known subgroup-specific variance to the evaluation labels; and **(ii)** model-error disparity, by scaling the model residuals within selected subgroups. Because both are injected at known magnitudes, we know the model disparity that RAFA should recover. On the nodules from LIDC dataset, subgroups are four random partitions, so baseline errors are equal in expectation; as a robustness check, we repeat the experiment with subgroups built by stratified matching, so that all four groups have the same distribution of nodule size and malignancy. In CheXpert, subgroups are *real*: race (four groups) and age (quartiles). Although CheXpert labels are binary, the model outputs continuous probabilities, so the signed error is continuous and additive Gaussian noise on the label gives a controlled test of the variance decomposition. Each experiment spans a $7 \times 7$ grid of label-noise and model-error disparities with each grid condition repeated $300$ times using independently sampled noise. The injected label-noise variance is known to the auditor here: this tests the adjustment itself, not the estimation of label noise, which is evaluated in Section 4 and, in a controlled form, in Appendix B.

**RAFA removes the label noise and keeps the model error.** Table 1 shows four representative regimes on the nodules. When neither label noise nor model disparity present (Q1), RAFA leaves everything unchanged. When only label noise disparity is present (Q2), the measured gap falls from $0.373$ to $0.075$. When only model disparity is present (Q3), RAFA leaves it unchanged ($0.907 \rightarrow 0.907$). When both are present (Q4), RAFA removes the label-induced component and recovers the known model gap ($1.119 \rightarrow 0.908$ vs. injected $0.907$). Across the full grid, the mean absolute difference between the adjusted gap and the known model gap is $0.003$ (random subgroups) and $0.004$ (matching subgroups). The radiographs reproduce this with real demographic subgroups: label-noise-only bias gaps of $0.62$ fall to $0.04$, injected model disparity is preserved, the frozen

Table 1: **Controlled validation on LIDC-IDRI.** RAFA removes disparity caused by unequal label noise (Q2), preserves injected model disparity (Q3), and separates the two in the mixed setting (Q4). Gaps are max–min subgroup error SD. The small Q1 gap (0.053) reflects finite-sample variation despite random subgroup assignment. "=" means no noise/error is added, while "≠" indicates controlled noise/error has been injected.

| Regime | Naive gap | RAFA-adjusted | Injected model gap |
|---|---|---|---|
| Q1 label=, model= (null) | 0.053 | 0.053 | 0.053 |
| Q2 label≠, model= (artifact) | 0.373 | 0.075 | 0.053 |
| Q3 label=, model≠ (model) | 0.907 | 0.907 | 0.907 |
| Q4 label≠, model≠ (mixed) | 1.119 | 0.908 | 0.907 |

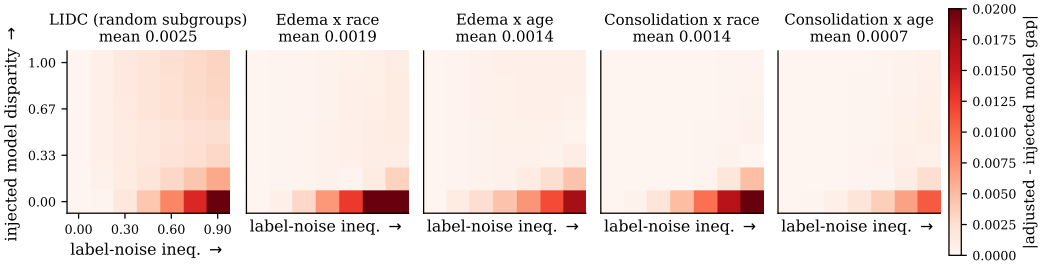



Figure 3: **Recovery across the controlled injection grids.** Each panel varies label-noise disparity (horizontal) and model-error disparity (vertical); color shows the absolute difference between the RAFA-adjusted gap and the known model gap. The left panel shows LIDC-IDRI with randomly assigned subgroups. The remaining panels show CheXpert with race and age subgroups for Edema and Consolidation. Recovery error remains close to zero across the grids. Note the color scale: its maximum is 0.02, while the injected gaps reach 0.9, so every cell is recovered to within about two percent of the largest gap. In particular, recovery stays accurate at the largest injected model disparities (top rows; Table 1, Q3–Q4).

model's pre-existing demographic gaps (0.007–0.033) are also preserved, and the mean difference between adjusted and known model gaps is 0.0007–0.0019 across the four finding–subgroup settings (Figure 3; full regimes in Appendix D). A correction that simply shrank every disparity would erase the model-error signal; RAFA does not. Sensitivity to a mis-estimated noise scale is asymmetric: under-estimation corrects incompletely but safely; over-estimation can over-correct, hiding the signal from the model's errors (Appendix C).

## 4 Case Study: Fetal-Weight Estimation

**Setting.** We now apply the full procedure to a real setting in which the target is not directly observed. We audit scan-time fetal-weight estimation from third-trimester ultrasound using 31,386 examinations from 6,513 pregnancies of 6,220 women across 17 hospitals (Mikołaj et al., 2025). The true fetal weight at scan time is unavailable. Instead, the evaluation label is obtained by projecting the recorded birth weight back to the scan gestational age using a population growth curve (Maršál et al., 1996). We define the scan-to-delivery (STD) interval as the number of days between the scan and birth. As STD grows, the back-projection spans a longer period during which an individual fetus may deviate from the population curve, so we expect the label to become noisier with increasing STD (R1). The audited deep-learning model uses ultrasound images, whereas the label error comes from fetal growth after the scan; these separate channels make zero covariance plausible (R3), and the audited predictions are out-of-sample, so the model could not fit the specific label errors of these cases. Neither property guarantees zero covariance, and we return to this below. We analyze STD because it directly controls the uncertainty of the evaluation label; any subgroup whose scan timing differs systematically inherits this source of apparent disparity. Demographic disparities in this cohort are studied in prior work (Elgebaly et al., 2026).

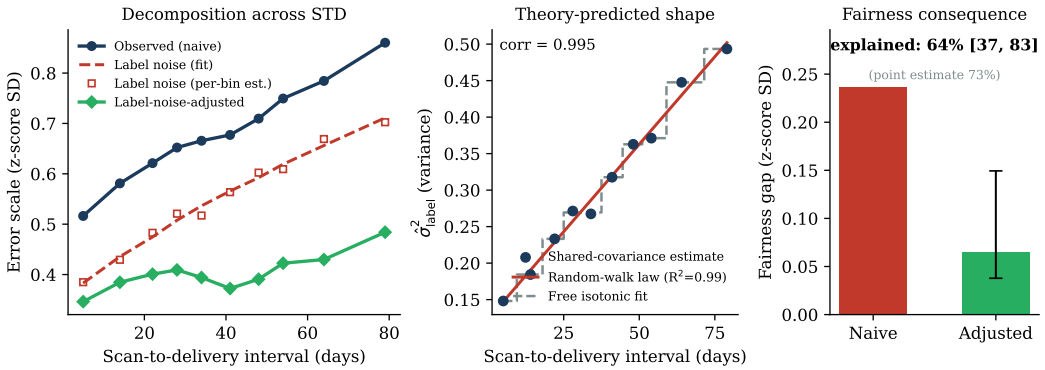



Figure 4: **End-to-end RAFA application to fetal-weight estimation. Left:** observed error increases with scan-to-delivery interval (STD); after accounting for estimated label noise, the increase is substantially reduced (ten equal-count STD bins; Table 2 gives four strata). **Middle:** covariance-based label-noise estimates follow the linear STD trend predicted by the random-walk model ($R^2 = 0.99$); an isotonic fit gives a similar shape (correlation 0.995). **Right:** the conventional disparity is 0.237 and the label-noise-adjusted disparity is 0.065. Bootstrap resampling over women gives a median fraction explained of 64% with a 95% CI of 37–83%. The interval is asymmetric around the point estimate because of the clip-induced positive bias near zero (Section 2).

Table 2: **Fetal-weight decomposition across STD strata.** Values are error SDs in $z$-score units. The adjustment is performed on the variance scale using Eq. (6); gaps in the text are calculated from unrounded values.

| STD stratum | $< 22$d | 22–38d | 38–54d | $> 54$d |
|---|---|---|---|---|
| Observed error (naive) | 0.573 | 0.650 | 0.708 | 0.809 |
| Label noise $\hat{\sigma}_{\text{label}}$ | 0.423 | 0.517 | 0.586 | 0.672 |
| Label-noise-adjusted | 0.386 | 0.394 | 0.398 | 0.451 |

**Estimating label noise from two predictors.** The label-noise variance is unknown here and must be estimated from the data. We use two different fetal-weight predictors: the audited deep-learning model and the Hadlock biometry formula (Hadlock et al., 1985). Both are evaluated against the same back-projected weight and therefore share the same error in the evaluation label. For predictor $i$, the measured error is $e_i = e_{m,i} - \epsilon$, where $e_{m,i}$ is the predictor's own error and $\epsilon$ is the shared label error. If the two predictors do not systematically make their own errors on the same cases, and neither predictor's error follows the label error, then the part of their measured errors that moves together comes from the shared label noise. Under these conditions:

$$\text{Cov}(e_1, e_2 \mid g) = \sigma^2_{\text{label}}(g). \tag{7}$$

We use the covariance between the two measured errors to estimate the label-noise variance in each subgroup. This construction keeps the estimate non-circular (R2): the model's own error variance cancels out of the covariance, so the estimate cannot absorb the model's own subgroup disparity. The main additional assumption is that the deep-learning model and Hadlock do not systematically fail on the same examinations. This is plausible because they use different mechanisms: the deep-learning model learns from ultrasound images, whereas Hadlock is a fixed formula on biometric measurements. It is not guaranteed: some examinations may be difficult for both, in which case the covariance would also contain shared prediction error and we would over-estimate the label noise. We therefore treat this estimator as valid under this explicit assumption and always report the conventional disparity alongside the adjusted one.

**The estimated noise grows with STD as predicted.** The back-projected label should get noisier the longer the interval between scan and birth, and the reason is cumulative: each day between scan and delivery, an individual fetus can deviate a little from the average growth curve, and these small deviations add up. If the daily deviations are roughly independent, their variances add, so the

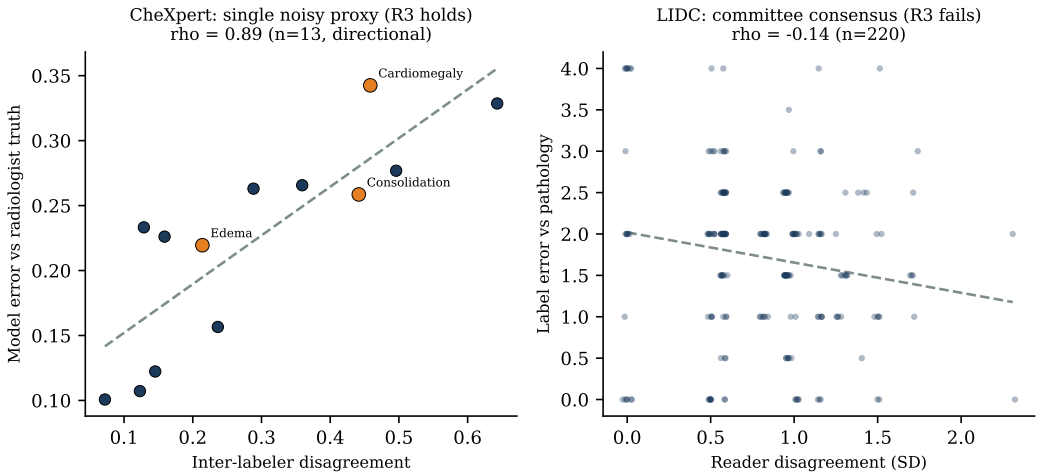



Figure 5: **When does disagreement reflect label quality? Left:** on CheXpert, disagreement between report-labeling systems is associated with model error against radiologist-adjudicated labels across findings (Spearman $\rho = 0.89$); highlighted points are the modeled findings. **Right:** on LIDC-IDRI, reader disagreement is weakly negatively associated with the consensus label's error against pathology-confirmed labels ($\rho = -0.14$; 220 nodules from 106 patients). The vertical axes differ by necessity: pathology adjudicates the label on nodules, while the radiologist set scores the model on radiographs; in both cases a valid noise signal should track error against the available truth.

variance of the accumulated deviation grows linearly with the number of days; this is the standard random-walk argument. The label-noise variance should therefore follow

$$\sigma^2_{\text{label}}(\text{STD}) = \sigma_0^2 + k\,\text{STD}. \tag{8}$$

The covariance-based estimates follow this line closely ($\sigma_0 = 0.351$, $k = 0.0048$, $R^2 = 0.99$). An isotonic regression, which assumes only that noise does not decrease with STD, gives nearly the same curve (correlation $0.995$ with the linear fit). Both fits use the same covariance estimates, so this is a consistency check on the expected shape, not an independent validation. For the adjustment we use the fitted line rather than the raw per-group estimates: with only two global parameters, the line cannot absorb per-group fluctuations of the model's own error.

**Adjusted disparity.** Table 2 shows the decomposition across four STD groups, and Figure 4 shows the trend at finer resolution. The conventional audit reports a gap of $0.237$ between the lowest- and highest-error groups. Subtracting the estimated label noise from every subgroup reduces it to $0.065$: on the full data, $73\%$ of the measured disparity is attributed to unequal label noise ($1 - \Delta_{\text{adj}}/\Delta_{\text{conv}}$). To propagate the uncertainty of the noise estimate itself, we use a cluster bootstrap: one woman can contribute several examinations and even several pregnancies, so we resample whole women and re-fit the noise law in every replicate. The median fraction explained is $64\%$ ($95\%$ CI 37–$83\%$). The result is stable when the law is fitted on one half of the data and applied to the other ($55\%$), and across three to six STD groups (63–$75\%$). A gap remains after adjustment, largest at the longest STD, so we do not conclude that model performance is equal across STD groups. What we conclude is that a substantial part of the disparity a conventional audit reports here comes from the labels, not the model.

## 5 When RAFA does not apply

RAFA needs a valid label-noise estimate (R1, R2) and separable errors (R3). We now look at natural settings: one where a noise signal carries real information, and three where the conditions fail. The supplementary material contains the full injection grids, the sensitivity analysis, the LIDC natural-label attempts, and the complete crowd-counting analysis.

**Disagreement can carry the signal, but not always.** On CheXpert we obtain three report-derived label sets (CheXpert, CheXbert, VisualCheXbert) and use their disagreement as a model-independent signal of label uncertainty. Disagreement varies across findings ($0.07$–$0.64$) and across race groups (range $0.244$), increases with age (range $0.150$), and barely varies across sex ($0.004$), a useful negative control. It is higher on examples which the labeler itself flags as uncertain ($0.689$ vs. $0.269$). Across the 13 findings, disagreement is strongly associated with the model's error against the radiologist labels (Spearman $\rho = 0.89$; Figure 5, left). With only 13 findings this is directional evidence, not a quantitative estimate of $\sigma^2_{\text{label}}(g)$, so we do not use it to compute a natural adjusted disparity. Whether disagreement carries this kind of information depends on how the labels are made: on LIDC, next, the same signal misleads.

**Committee disagreement does not measure consensus error.** On LIDC-IDRI, the evaluation label is a committee consensus, and it is still a noisy estimate of the truth. The problem is the estimate: reader disagreement measures the spread of the committee around its consensus, not the error of the consensus itself, so it is the wrong quantity to plug in as $\sigma^2_{\text{label}}(g)$. Indeed, the disagreement-based variance estimate exceeds the model's observed error variance in every subgroup, so the clip in Eq. (6) binds. Against pathology-confirmed labels, reader disagreement is even weakly negatively correlated with consensus-label error (Spearman $\rho = -0.14$, $p = 0.03$; Figure 5, right). Reader disagreement is not a suitable estimate of $\sigma^2_{\text{label}}(g)$ here. Further constructions are reported in Appendix E.

**A single binary label is not enough.** With one binary evaluation label per image, as in the natural CheXpert audit, the variance of the observed label within a subgroup is $p(1-p)$, where $p$ is the fraction of positive labels. That quantity is fixed by the labels being audited: it mixes the true prevalence of the finding with label error, and the two cannot be separated without extra information such as repeated labels for the same image. It is therefore not the label-noise variance RAFA needs.

**Shared task difficulty: crowd counting.** The two previous violations come from how the labels are generated. Crowd counting shows a violation that comes from the task itself. The task is to predict the number of people in an image, and the evaluation label is a manual count: an annotator places a dot on every visible head. As in the fetal audit, label quality depends on a measurable parameter: annotating a head that spans hundreds of pixels is easy, annotating one a few pixels wide is hard, so label noise grows as heads get more distant. The difference is where the difficulty lives. In the fetal audit the label error comes from fetal growth after the scan, which the image-based model never sees. In crowd counting the same visual difficulty hits the annotator and every model alike. Section 4 flagged shared difficulty as the residual risk of the fetal audit; crowd counting shows what happens when it is fully realized. Heads near the camera span tens to hundreds of pixels; heads far away only a few. Following Zhang et al. (2016); Shi et al. (2019), we measure head size by the mean distance to the three nearest annotated heads and form far, mid, and near subgroups by percentiles (details in Appendix F). On two benchmarks, SHHA (Zhang et al., 2016) and UCF-QNRF (Idrees et al., 2018), we evaluate five models with different training objectives: CSRNet (Li et al., 2018), DM-Count (Wang et al., 2020a), P2PNet (Song et al., 2021), PET (Liu et al., 2023), and CLIP-EBC (Ma et al., 2025). If the models' own errors were uncorrelated, as the two-predictor construction of Eq. (7) assumes, every off-diagonal covariance would estimate $\sigma^2_{\text{label}}(g)$. Instead, the model errors co-fail: in every subgroup one dominant principal component with positive loadings explains 73–87% of the error variance, and mean pairwise correlations are $0.83/0.72/0.67$ (far/mid/near) on SHHA and $0.71/0.73/0.72$ on UCF-QNRF. Even the smallest pairwise covariance, the most favorable estimate, removes 27–84% of each model's error variance across the 30 model–subgroup cells, and the mean covariance drives Eq. (6) to the clip for the best model in five of six subgroups. The estimate is dominated by shared model failure, not label noise, breaking the assumption behind Eq. (7) (R2). The labeler route fails too (R3): for distant heads the dominant label error is a systematic miss below the resolution limit, so $\mathbb{E}[\epsilon \mid g] < 0$; two independent annotators under-count in the same way, which a variance-based disagreement estimate cannot see, and every model trained on such labels learns to under-count the far group, so $e_m$ tracks $\epsilon$ and $\text{Cov}(e_m, \epsilon \mid g) > 0$. The two violations reinforce each other: R3 makes the observed model error look smaller than it is, R2 makes the estimated label noise look larger than it is. Put together, they push the adjusted error toward zero exactly where the models are worst, absorbing model errors as label noise.

## 6 DISCUSSION

**Arguing the assumptions in practice.** R2 and R3 cannot be tested from the audit data, so they must be argued from how the labels and the noise estimate are produced. Our applications suggest four practical checks; none is sufficient alone. First, check where the noise estimate comes from. If it never uses the audited model's predictions, it cannot copy the model's error pattern directly; the CheXpert disagreement signal, computed from report text alone, is one example, although it is not precise enough to use for the adjustment (Section 5). This rules out direct circularity only: the estimate and the model can still move together through a shared cause, when the same cases are hard for the labelers and the model, or when the model has inherited errors from the noisy labels it was trained on. Those channels are what R3 and the remaining checks address. Second, give the noise estimate few degrees of freedom. In the fetal study, the estimated noise is a straight line in STD with two parameters fitted on the whole dataset; a two-parameter line cannot reproduce an arbitrary pattern across subgroups, so it cannot quietly absorb the model's own disparity. Third, run a negative control: pick a grouping where label noise has no reason to differ, and check that the estimate is flat there. On CheXpert, disagreement differs across race and age groups but not between sexes, as expected. Fourth, compare the size of the estimate with what is physically plausible; shared difficulty inflates it beyond any plausible labeling error. In crowd counting, the estimated annotation error reaches tens of heads per image (Appendix F), far more than annotators plausibly miss, which exposes the estimate as shared model failure. None of these checks proves the assumptions, but each would likely catch a serious violation.

**Be conservative.** Mis-estimating the noise scale fails asymmetrically (Appendix C): under-estimation corrects incompletely but never manufactures an artificial disparity, while over-estimation can over-correct and flip the conclusion. When the noise magnitude is uncertain, use a conservative estimate and always report the conventional disparity alongside the adjusted one.

**Limitations and future work.** RAFA applies to disparities defined on continuous signed errors, where the variance decomposition of Eq. (3) holds. Metrics such as false-negative rate and AUC do not decompose the same way; extending label-noise adjustment to them is a natural next step. A second direction is building noise estimators for new label types: the adjustment itself is validated here, and each new application needs an estimate constructed from how its own labels are made.

**Conclusion.** Fairness audits implicitly assume equally accurate labels for every subgroup. When label noise differs across subgroups, part of the reported disparity comes from the labels rather than the model. RAFA separates the two on the variance scale: in controlled experiments on real predictions and real demographic subgroups it removes injected label-noise disparity while preserving known model disparity, and in a full clinical application it attributes most of an apparent fairness gap to unequal label noise. It comes with explicit validity conditions and practical ways to check them. Because RAFA changes the evaluation and not the model, it complements any mitigation method.

### REPRODUCIBILITY STATEMENT

Code for all experiments, including the simulation, controlled injection grids, sensitivity analyses, fetal decomposition and bootstrap, crowd-counting analysis, and scripts used to produce the reported figures and tables, will be released on publication. All controlled experiments and bootstrap analyses use fixed random seeds. LIDC-IDRI, CheXpert, SHHA, and UCF-QNRF are publicly available. The fetal cohort is a restricted clinical dataset and cannot be released; we will provide the analysis code used for that study.

### AI USE STATEMENT

Generative AI tools were used during the preparation of this paper to improve the language and grammar and to provide feedback on clarity and organization. All experimental design choices, analyses, results, interpretations, and final scientific claims are entirely the responsibility of the authors.

## REFERENCES

J. Adebayo, Melissa Hall, Bowen Yu, and Bobbie Chern. Quantifying and mitigating the impact of label errors on model disparity metrics. *ArXiv*, abs/2310.02533, 2023. URL https://api.semanticscholar.org/CorpusId:259298198.

Olalekan Joseph Akintande, Siavash Arjomand Bigdeli, and Aasa Feragen. Medicine after death: Xai and algorithmic fairness under label bias. In *Fourth European Workshop on Algorithmic Fairness*, pp. 171–186. ML Research Press, 2025.

Junaid Ali, Preethi Lahoti, and Krishna P. Gummadi. Accounting for model uncertainty in algorithmic discrimination. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 336–345, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462630. URL https://doi.org/10.1145/3461702.3462630.

Samuel G. Armato III, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011. doi: 10.1118/1.3528204.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge, MA, 2023. ISBN 9780262048613.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, 1993. ISBN 9780412042317.

Aya Elgebaly, Joris Fournel, Benjamin Laine Jønch Jurgensen, Kamil Mikolaj, Anders Nymark Christensen, Martin Grønnebæk Tolsgaard, Claes Ladefoged, and Aasa Feragen. Intersectional disentangling of temporal and acquisition bias in fetal ultrasound. In *Joint FAIMI, BRIDGE, and EPIMI Workshop at MICCAI 2026*, 2026. URL https://openreview.net/forum?id=SZ4ASOImb0.

Riccardo Fogliato, Alexandra Chouldechova, and Max G'sell. Fairness evaluation in presence of biased noisy labels. In *International conference on artificial intelligence and statistics*, pp. 2325–2336. PMLR, 2020.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894.

Wayne A. Fuller. *Measurement Error Models*. John Wiley & Sons, New York, 1987. ISBN 9780470316665.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis T. Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022. doi: 10.1016/S2589-7500(22)00063-2.

Frank P. Hadlock, R. B. Harrist, Ralph S. Sharman, Russell L. Deter, and Seung K. Park. Estimation of fetal weight with the use of head, body, and femur measurements—a prospective study. *American Journal of Obstetrics and Gynecology*, 151(3):333–337, 1985. doi: 10.1016/0002-9378(85)90298-4.

Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Computer Vision – ECCV 2018*, pp. 544–559, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01216-8. doi: 10.1007/978-3-030-01216-8_33.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2):138–146, 2024.

S. Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. Uncertainty as a fairness measure. *J. Artif. Intell. Res.*, 81:307–335, 2023. URL https://api.semanticscholar.org/CorpusId:266359296.

Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100, 2018. doi: 10.1109/CVPR.2018.00120.

Yunyi Li, Maria De-Arteaga, and M. Saar-Tsechansky. Mitigating label bias via decoupled confident learning. *ArXiv*, abs/2307.08945, 2023. URL https://api.semanticscholar.org/CorpusId:259951486.

Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongliang Liu. Point-query quadtree for crowd counting, localization, and more. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1676–1685, 2023. doi: 10.1109/ICCV51070.2023.00161.

Yiming Ma, Victor Sanchez, and Tanaya Guha. Clip-ebc: Clip can count accurately through enhanced blockwise classification. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209839.

Karel Maršál, Per-Henrik Persson, Torbjörn Larsen, Hans Lilja, Anders Selbing, and Bengt Sultan. Intrauterine growth curves based on ultrasonically estimated foetal weights. *Acta Paediatrica*, 85 (7):843–848, 1996. doi: 10.1111/j.1651-2227.1996.tb14164.x.

Vishwali Mhasawade, Alexander D'Amour, and Stephen R. Pfohl. A causal perspective on label bias. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. URL https://api.semanticscholar.org/CorpusId:270286491.

Kamil W. Mikołaj, Anders N. Christensen, Caroline A. Taksøe-Vester, et al. Predicting abnormal fetal growth using deep learning. *npj Digital Medicine*, 8:318, 2025. doi: 10.1038/s41746-025-01704-0.

Abhilash Mishra and Yash Gorana. Who decides if ai is fair? the labels problem in algorithmic auditing. *ArXiv*, abs/2111.08723, 2021. URL https://api.semanticscholar.org/CorpusId:244270509.

Aditya Parikh, Sneha Das, and Aasa Feragen. Investigating label bias and representational sources of age-related disparities in medical segmentation. *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2025. URL https://api.semanticscholar.org/CorpusId:282739070.

Aditya Parikh, Stella Frank, Snehayan Das, and A. Feragen. Towards fairness under label bias in image segmentation: Impact, measurement and mitigation. *ArXiv*, abs/2605.06891, 2026. URL https://api.semanticscholar.org/CorpusId:288110813.

Srinivasan H. Sengamedu and Hien Pham. Fairlabel: Correcting bias in labels. *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 262–271, 2023. URL https://api.semanticscholar.org/CorpusId:264833076.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021. doi: 10.1038/s41591-021-01595-0.

Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7271–7280, 2019. doi: 10.1109/CVPR.2019.00745.

Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3345–3354, 2021. doi: 10.1109/ICCV48922.2021.00335.

E. A. M. Stanley, R. Mehta, M. Roschewitz, N. D. Forkert, and B. Glocker. Exploring the interplay of label bias with subgroup size and separability: A case study in mammographic density classification. In *Fairness of AI in Medical Imaging*, volume 15976 of *Lecture Notes in Computer Science*, Cham, 2026. Springer. doi: 10.1007/978-3-032-05870-6_8. URL https://doi.org/10.1007/978-3-032-05870-6_8.

Anique Tahir, Lu Cheng, and Huan Liu. Fairness through aleatoric uncertainty. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023. URL https://api.semanticscholar.org/CorpusId:258041323.

John R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. University Science Books, Sausalito, CA, 2nd edition, 1997. ISBN 9780935702750.

Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1595–1607. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/118bd558033a1016fcc82560c65cca5f-Paper.pdf.

Jialu Wang, Yang Liu, and Caleb C. Levy. Fair classification with group-dependent label noise. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2020b. URL https://api.semanticscholar.org/CorpusId:226227049.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 589–597, June 2016.

# Supplementary Material

## A LIDC CONTROLLED-INJECTION GRID

The main-text LIDC experiment independently varies label-noise disparity and model-error disparity on real malignancy predictions. Subgroups are constructed either by random assignment or by stratified matching, which balances the four groups on nodule size and malignancy. For each construction, we evaluate seven levels of label-noise inequality crossed with seven levels of model-error inequality, giving a $7 \times 7$ grid. Each cell is repeated over 300 random draws.

Label-noise disparity is introduced by adding mean-zero noise with known subgroup-specific variance to the evaluation labels. Model-error disparity is introduced separately by scaling the model residuals within selected subgroups. The injected label-noise variance is supplied directly to RAFA in this experiment, so the analysis isolates the adjustment in Eq. (6) from the separate problem of estimating label noise.

For each grid cell, we compare the RAFA-adjusted disparity with the model disparity known from the controlled injection. We define the recovery error as

$$E_{\mathrm{rec}} = \left|\Delta_{\mathrm{adj}} - \Delta_{\mathrm{model}}\right|, \tag{9}$$

where $\Delta_{\mathrm{adj}}$ is the disparity after RAFA adjustment and $\Delta_{\mathrm{model}}$ is the known model disparity in that controlled setting.

Across the full grid, the mean recovery error is $0.003$ under random subgroup assignment and $0.004$ after matching. The low recovery error across increasing levels of label noise shows that the adjustment tracks the model component rather than the total observed disparity.

## B ESTIMATING THE INJECTED NOISE: CORRUPTION RECOVERY

The grid experiments supply the injected label-noise variance directly to RAFA. Here we instead make RAFA estimate it. Using the real LIDC predictions with identical model behavior across four synthetic subgroups, we corrupt the evaluation labels with Gaussian noise of known per-group scale ($\sigma = 0, 0.3, 0.6, 0.9$) and estimate the label-noise variance from repeated corrupted measurements of each label (pooled within-item variance), over $300$ seeds.

The estimated noise scales match the injected ones (correlation $1.000$, mean absolute error $0.035$). The manufactured gap of $0.370$ falls to $0.094$ with the estimated noise ($74.6\%$ removed), against $0.077$ with the oracle noise ($79.2\%$). This is a best-case setting: the noise is Gaussian and repeated measurements of each label are available. It shows the estimation-plus-adjustment pipeline is correct end to end, not that estimation is easy in general; the natural-label attempts in Appendix E show the opposite case.

## C SENSITIVITY TO LABEL-NOISE MIS-ESTIMATION

RAFA depends on the estimated subgroup label-noise scale $\sigma_{\hat{\text{label}}}(g)$. We therefore test how systematic and random errors in this estimate affect the resulting disparity.

**Systematic mis-estimation.** We multiply the true label-noise scale by a common factor before applying RAFA. The fraction of the manufactured disparity removed is

| Scale factor | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 |
|---|---|---|---|---|---|
| Gap removed | 22% | 51% | 80% | 14% | −147% |

Under-estimation leaves part of the label-induced disparity uncorrected. Over-estimation is more problematic: excessive subtraction can increase the adjusted disparity rather than remove it. Negative removal therefore indicates that the adjusted gap is larger than the original gap.

**Random estimation error.** We additionally perturb $\sigma_{\hat{\text{label}}}(g)$ with mean-zero noise of standard deviation $\tau$. The fraction of the manufactured gap removed decreases as the estimation error increases:

| $\tau$ | 0 | 0.1 | 0.2 | 0.3 |
|---|---|---|---|---|
| Gap removed | 80% | 58% | 17% | −29% |

These experiments show an asymmetric failure mode. Conservative under-estimation produces incomplete adjustment, whereas substantial over-estimation can over-correct the observed disparity. When the label-noise magnitude is uncertain, this motivates preferring a conservative estimate and reporting the conventional disparity alongside the RAFA-adjusted result.

## D CHEXPERT CONTROLLED-INJECTION GRIDS

We repeat the controlled-injection experiment on chest-radiograph predictions, using real demographic subgroup definitions rather than randomly constructed groups. We evaluate Edema and Consolidation predictions across race groups and age quartiles.

As in the LIDC experiment, label-noise disparity and model-error disparity are controlled independently over a $7 \times 7$ grid. The null settings preserve the baseline demographic differences already present in the frozen model rather than forcing the subgroup errors to be identical.

Across the four task–subgroup combinations, the null disparities range from $0.007$ to $0.033$ and remain unchanged after adjustment. In the all-model regimes, RAFA preserves the injected model disparity. In the all-label-noise regimes, observed gaps of approximately $0.62$ decrease to approximately $0.04$ after adjustment. Across the full grids, the mean absolute difference between the adjusted and known model disparities ranges from $0.0007$ to $0.0019$.

These results reproduce the LIDC pattern with naturally occurring demographic subgroups: RAFA removes the controlled label-induced component while preserving the model component.

## E LIDC NATURAL-LABEL ATTEMPTS

Beyond the committee analysis in Section 5, we tested four constructions on the natural LIDC labels. A reader used as the predictor is circular by construction and the adjusted gap collapses to zero. For trained ResNet classifiers, the disagreement-based noise estimate exceeds the observed error variance in every size bin, so the clip binds. A strong segmentation model (out-of-fold Dice 0.861) makes the degeneracy worse: its error (0.139) is smaller than the inter-reader disagreement (0.239) in every bin. Finally, the pathology check of Figure 5 (right) shows that reader disagreement does not track where the consensus label is wrong. All four fail the same way: committee spread is not the additive noise inside the model's measured error.

## F ERROR ANALYSIS IN CROWD COUNTING

We document the results for the crowd counting case. Figures 6 and 9 demonstrate how the labels are split into different subgroups according to the total headsize distribution in each dataset, into far, mid and near subgroups. Figures 7 and 10 illustrate how strongly correlated the models' errors are. The naive and adjusted bias are presented in Figures 8 and 11, as well as in Tables 3 and 4.

### F.1 RESULTS ON THE SHANGHAITECH–A DATASET




(a) IMG_38, 717 heads. (b) IMG_182, 199 heads.

Figure 6: Head-size groups on two ShanghaiTech A test images. Each annotated head is coloured by its group, defined from the mean distance to its three nearest annotated heads: far (<12 px), mid (12–25 px), near (≥25 px). Far heads dominate the distant, dense regions; near heads the foreground.

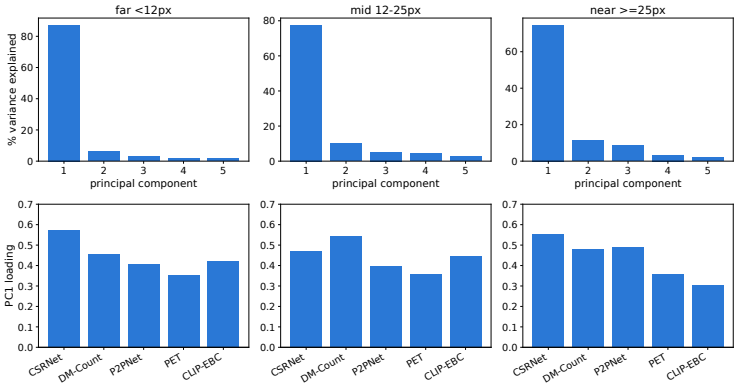



Figure 7: PCA analysis of the $5 \times 5$ model error covariance matrix on the SHHA dataset, with three head-size subgroups: far, mid, and near. PC1 explains 75–87% of the total error variance in all three cases, with positive loadings, showing that the models' failures are correlated.

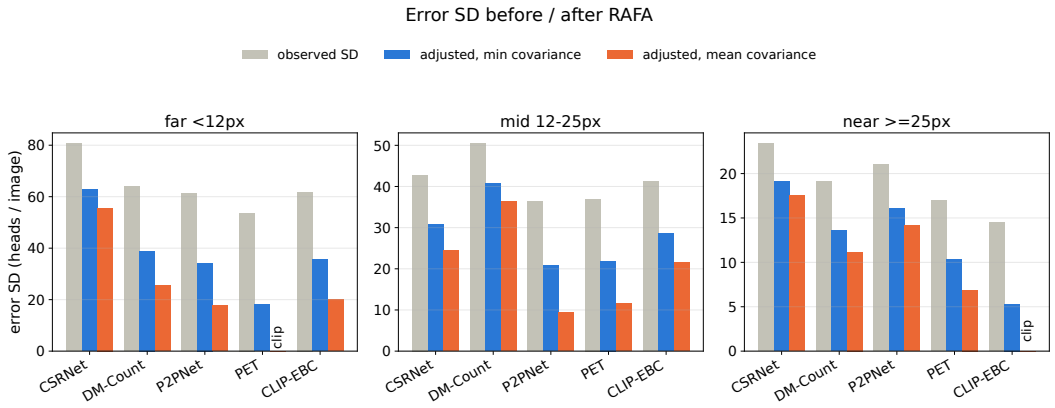
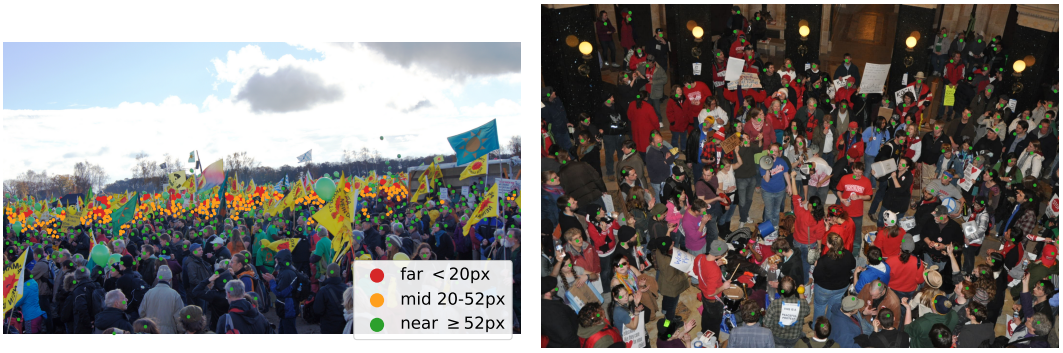



Figure 8: RAFA-adjusted SD values for the SHHA dataset across the five models, with the minimum off-diagonal estimate (blue, $\min \hat{\sigma_{\text{label}}}(\text{far/mid/near}) = 50.6/29.8/13.5$) and the mean off-diagonal estimate (orange, $\text{mean } \hat{\sigma_{\text{label}}}(\text{far/mid/near}) = 58.5/35.1/15.5$) of the label noise.

Table 3: Disparity across head-size groups before and after RAFA on ShanghaiTech A. $\Delta = \max_g s(g) - \min_g s(g)$ over the far/mid/near error SDs (heads per image); the fraction explained is $1 - \Delta_{\text{adj}}/\Delta_{\text{conv}}$. $\hat{\sigma_{\text{label}}}^2(g)$ is the minimum or the mean off-diagonal error covariance of the group.

| | | Min. covariance | | Mean covariance | |
|---|---|---|---|---|---|
| Model | $\Delta_{\text{conv}}$ | $\Delta_{\text{adj}}$ | Explained | $\Delta_{\text{adj}}$ | Explained |
| CSRNet | 57.3 | 43.8 | 24% | 38.1 | 34% |
| DM-Count | 44.7 | 27.2 | 39% | 25.1 | 44% |
| P2PNet | 40.1 | 18.2 | 55% | 8.2 | 80% |
| PET | 36.7 | 11.5 | 69% | 11.5 | 69% |
| CLIP-EBC | 47.3 | 30.2 | 36% | 21.6 | 54% |

## F.2 Results on the UCF–QNRF Dataset




(a) img_0211, 434 heads. (b) img_0170, 148 heads.

Figure 9: Head-size groups on two UCF-QNRF test images. Each annotated head is coloured by its group, defined from the mean distance to its three nearest annotated heads: far (<20 px), mid (20–52 px), near (≥52 px). Far heads dominate the distant, dense regions; near heads the foreground.

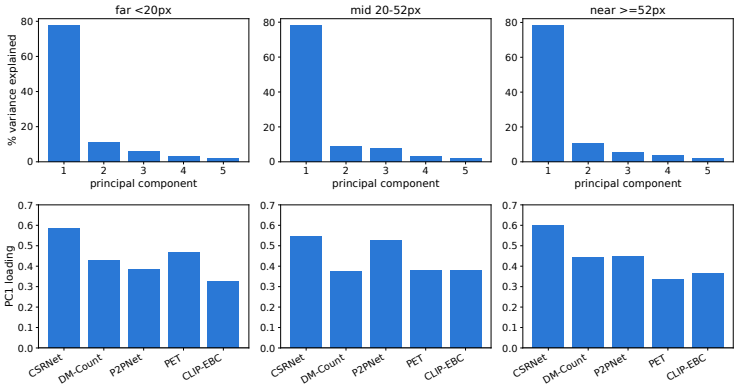



Figure 10: PCA analysis of the $5 \times 5$ model error covariance matrix on the UCF-QNRF dataset, with three head-size subgroups: far, mid, and near. PC1 explains 73% of the total error variance in all three cases, with positive loadings, showing that the models' failures are correlated.

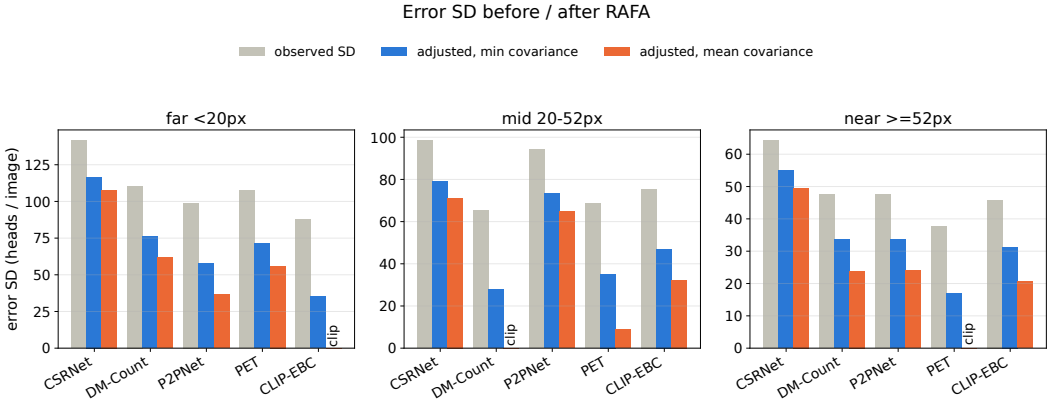



Figure 11: RAFA-adjusted SD values for the UCF-QNRF dataset across the five models, with the minimum off-diagonal estimate (blue, $\min \hat{\sigma_{\text{label}}}(\text{far/mid/near}) = 80.5/59.0/33.5$) and the mean off-diagonal estimate (orange, $\text{mean } \hat{\sigma_{\text{label}}}(\text{far/mid/near}) = 91.9/68.1/41.0$) of the label noise.

Table 4: Disparity across head-size groups before and after RAFA on UCF-QNRF (percentile edges). $\Delta = \max_g s(g) - \min_g s(g)$ over the far/mid/near error SDs (heads per image); the fraction explained is $1 - \Delta_{\text{adj}}/\Delta_{\text{conv}}$. $\hat{\sigma_{\text{label}}}^2(g)$ is the minimum or the mean off-diagonal error covariance of the group.

| | | Min. covariance | | Mean covariance | |
|---|---|---|---|---|---|
| Model | $\Delta_{\text{conv}}$ | $\Delta_{\text{adj}}$ | Explained | $\Delta_{\text{adj}}$ | Explained |
| CSRNet | 77.3 | 61.7 | 20% | 58.2 | 25% |
| DM-Count | 63.2 | 48.1 | 24% | 61.6 | 3% |
| P2PNet | 51.4 | 39.6 | 23% | 40.8 | 21% |
| PET | 70.1 | 54.5 | 22% | 56.0 | 20% |
| CLIP-EBC | 42.0 | 15.7 | 63% | 32.4 | 23% |